

# Outage Effective Capacity of Buffer-Aided Diamond Relay Systems Using HARQ with Incremental Redundancy

Deli Qiao

## Abstract

In this paper, transmission over buffer-aided diamond relay systems under statistical quality of service (QoS) constraints is studied. The statistical QoS constraints are imposed as limitations on delay violation probabilities. In the absence of channel state information (CSI) at the transmitter, truncated hybrid automatic repeat request-incremental redundancy (HARQ-IR) is incorporated to make better use of the wireless channel and the resources for each communication link. The packets that cannot be successfully received upon the maximum number of transmissions will be removed from buffer, i.e., outage occurs. The *outage effective capacity* of a communication link is defined as the maximum constant arrival rate to the source that can be supported by the *goodput* departure processes, i.e., the departure that can be successfully received by the receiver. Then, the outage effective capacity for the buffer-aided diamond relay system is obtained for HARQ-IR incorporated transmission strategy under the *end-to-end* delay constraints. In comparison with the DF protocol with perfect CSI at the transmitters, it is shown that HARQ-IR can achieve superior performance when the SNR levels at the relay are not so large or when the delay constraints are stringent.

## I. INTRODUCTION

In wireless systems, the power of the received signal fluctuates randomly over time due to mobility, changing environment, and multipath fading caused by the constructive and destructive

This work has been supported in part by the National Natural Science Foundation of China (61671205) and the Shanghai Sailing Program (16YF1402600).

Part of this work has been submitted to the 2017 IEEE International Conference on Communications (ICC) [24].

D. Qiao is with the School of Information Science&Technology, East China Normal University, Shanghai, China 200241. Email: dlqiao@ce.ecnu.edu.cn.

superimposition of the multipath signal components [25]. These random changes in the received signal strength lead to variations in the instantaneous data rates that can be supported by the channel, which may result in transmission errors in deep fading. Hybrid automatic repeat request (HARQ) protocols have been proposed to enhance the wireless systems performance. Generally, the receiver sends either an acknowledgement (ACK) or negative ACK (NACK) to the transmitter depending on whether the data packet is correctly received or not. The transmitter can decide either to send the next packet or retransmit the same packet upon reception of ACK or NACK, respectively [1]. The performance of ARQ protocols has been extensively studied in literature (see e.g., [2]-[5] and references therein).

Also, relay channels can be viewed as one of the basic building blocks of wireless systems. Information-theoretic analysis of relay channels has been the research forefront for decades, and has shown the performance improvement in terms of throughput and diversity (see, e.g., [6]-[13]). For instance, the authors have considered different relaying strategies in [7], and showed that considerable cooperative diversity can be achieved with the relaying schemes. The authors have derived the expressions for the outage probability and throughput for HARQ protocols in relay channels in [8]. Of particular interest is the *diamond relay system* in which the communication between a disconnected source and destination is achieved via the help of two or more intermediate relay nodes. The authors have analyzed the capacity bounds for the full-duplex relays with additive white Gaussian channels in [9], while different transmission strategies and achievable rates in half-duplex Gaussian diamond relay channel have been investigated in [10]. The authors have characterized the outage probability and throughput of HARQ protocols with relay selection for the multirelay channels in [13]. More recently, buffer-aided relaying in which the relays are equipped with buffers have been shown to further improve the performance of relay systems [14], [15]. Design and analysis of buffer-aided relay systems have attracted much interest recently [15].

Generally, information theoretic analysis do not take into account the buffer/queue limitations. In present wireless systems, diverse quality of service (QoS) requirements are driven by the exponential growth of wireless multimedia traffic that is generated by smartphones, tablets, servers, social networking tools and video sharing sites. In multimedia applications involving e.g., voice over IP (VoIP), streaming video, and interactive video, certain QoS limitations in terms of buffer/delay constraints are imposed so that target levels of performance and quality

can be provided to the users. The concept of effective capacity [16] has been incorporated to characterize the maximum constant arrival rate under statistical delay constraints. In case of point-to-point links, there have been some related works investigating the HARQ protocols of wireless channels under statistical QoS constraints recently [17]-[22]. For instance, in [17], we have analyzed the energy efficiency of fixed rate transmissions under statistical QoS constraints with a simple Type-I HARQ (HARQ-T1) protocol. In this work, we assumed that no outage occurs, i.e., retransmissions are triggered as long as long the receiver does not receive the packet. In [19], the author has analyzed the performance of HARQ with incremental redundancy (HARQ-IR), and showed that with stringent QoS constraints, HARQ-IR can outperform the adaptive transmission system. In [20], the authors have investigated fixed rate transmissions with HARQ protocols, and obtained the closed-form expression for the effective capacity of HARQ-IR only for loose QoS constraints. In [21], the authors have characterized the effective capacity of different HARQ protocols with limited number of transmissions, or deadline of the packets. Outage occurs when the packet is dropped from the buffer while the receiver does not correctly receive the packet. However, the effective capacity obtained does not specify the average throughput that can be correctly received at the receiver. In [22], the authors have considered the goodput of various HARQ protocols, and proposed a general framework to express effective capacity of HARQ protocols based on a random walk model and recurrence relation formulation. In this paper, we present a study on the buffer-aided diamond relay systems with HARQ-IR under statistical QoS constraints, in the form of limitations on the delay violation probabilities.

In this work, we assume that the channel state information (CSI) is absent at the transmitters for the links. We first define the *outage effective capacity* as the maximum constant arrival rate that can be supported by the departure processes correctly received at the receiver while satisfying the statistical QoS constraints for a communication link. We show that there is an optimal fixed transmission rate with HARQ-IR scheme. We also demonstrate that the outage effective capacity approaches to the throughput of the link as the delay constraints vanish. We then consider full-duplex decode-and-forward (DF) relays, and assume that the source sends the common information to the relays, which cooperatively deliver the same message to the destination. The relays adopt the Alamouti scheme to enhance the information delivery to the destination. With the proposed HARQ-IR scheme, we derive the outage effective capacity of the buffer-aided diamond relay system and the associated outage probability. For comparison,

we also consider the typical DF protocol [12] in case of perfect CSI at the transmitter and the receiver for all links, where the common information is sent by the source at the minimum rate of the source-relay links and distributed beamforming is performed at the relays. The contributions of this work can be summarized as follows:

- 1) We obtain the outage effective capacity of the goodput processes of a communication link for the HARQ protocols following the spectral radius method, and prove that the limiting behavior of the resulting expression coincides with several well-known results, such as the throughput of HARQ protocols without delay constraints and the effective capacity of HARQ-T1 protocol with unlimited number of transmissions;
- 2) We propose a HARQ-IR based transmission scheme for the buffer-aided diamond relay systems with perfect CSI at the receiver only for each link, and characterize the outage effective capacity of the proposed scheme under the statistical delay constraints;
- 3) Through numerical evaluations, we demonstrate the superiority of the proposed scheme with respect to the DF protocol with perfect CSI at the transmitter and receiver of each link when the SNR at the relay is relatively small or when the delay constraints are relatively stringent.

The rest of this paper is organized as follows. Section II introduces preliminaries on the diamond relay channel model, and reviews the HARQ-IR operations. In Section III, we briefly discuss the statistical delay constraints and define the outage effective capacity for one-hop links. Section IV discusses the effective capacity analysis method for two-hop links, and characterize the outage effective capacity of the buffer-aided diamond relay systems. Numerical results are provided in Section V. Finally, Section VI concludes this paper.

## II. PRELIMINARIES

### *A. System Model*

We consider a buffer-aided diamond relay communication link as depicted in Fig. 1. The source sends information to the destination via the help of two parallel relays. We assume that there is no direct link between the source and the destination. Also, there is no link between the relays. In this model, there are buffers of infinite size at both the source and relays. In this work, we assume full-duplex relay such that transmission and reception can be performed simultaneously.

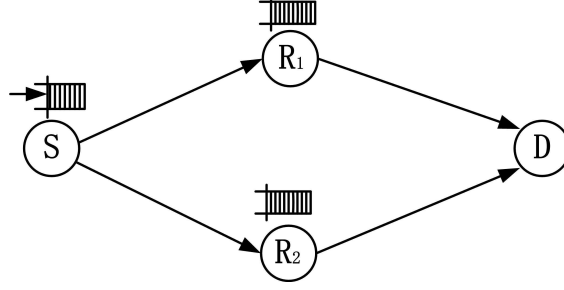


Fig. 1. The diamond-relay system model.

The discrete-time input and output relationships in the  $i$ th symbol duration are given by

$$Y_{r_j}[i] = g_{sr_j}[i]X_s[i] + n_{r_j}[i], \quad j = 1, 2, \quad (1)$$

$$Y_d[i] = g_{r_1d}[i]X_{r_1}[i] + g_{r_2d}[i]X_{r_2}[i] + n_d[i], \quad (2)$$

where  $X_k$  for  $k \in \{s, r_1, r_2\}$  denote the input signal from the source **S** and the relay **R<sub>j</sub>**,  $j = 1, 2$ , respectively. The inputs are subject to individual average energy constraints  $\mathbb{E}\{|X_k|^2\} \leq \bar{P}_k/B$ ,  $k \in \{s, r_1, r_2\}$ , where  $B$  is the bandwidth.  $Y_{r_j}, Y_d$  represent the received signal at the relay **R<sub>j</sub>** and the destination **D**, respectively. We assume that the fading coefficients  $g_{sr_j}, g_{r_jd}$  are jointly stationary and ergodic discrete-time processes, and we denote the magnitude-square of the fading coefficients by  $z_{sr_j}[i] = |g_{sr_j}[i]|^2$  and  $z_{r_jd}[i] = |g_{r_jd}[i]|^2$ . Denote  $\mathbf{z} = (z_{sr_1}, z_{sr_2}, z_{r_1d}, z_{r_2d})$ . Assuming that there are  $B$  complex symbols per second, we can easily see that the symbol energy constraint of  $\bar{P}_k/B$  implies that the channel input has a power constraint of  $\bar{P}_k$ . Above, in the channel input-output relationships, the noise component  $n_k[i]$  is a zero-mean, circularly symmetric, complex Gaussian random variable with variance  $\mathbb{E}\{|n_k[i]|^2\} = N_0$  for  $k \in \{r_1, r_2, d\}$ . The additive Gaussian noise samples  $\{n_k[i]\}$  are assumed to form an independent and identically distributed (i.i.d.) sequence. We denote the signal-to-noise ratio at source as  $\text{SNR}_s = \frac{\bar{P}_s}{N_0B}$ , and at relays as  $\text{SNR}_{r_j} = \frac{\bar{P}_{r_j}}{N_0B}$ ,  $j = 1, 2$ .

### B. HARQ-IR

Consider a link composed of one transmitter and one receiver under block fading in which the fading stays constant for a block of  $T$  seconds and changes independently from one block to another. We assume that a packet of  $L$  bits is intended to be transmitted over the wireless channel

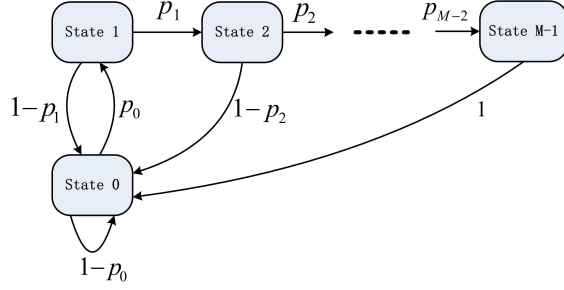


Fig. 2. The state transition model.

in each frame. Specifically, after each successful transmission, the transmitter attempts to send  $L$  bits in the next frame. So the fixed transmission rate is termed as  $L$  bits/block. We assume that upon successful reception of the packet, the receiver sends an ACK to the transmitter, and the packet can be removed from the buffer. If a decoding failure occurs, the receiver sends a NACK to the transmitter and requests another round of retransmissions for the packet if the maximum number of transmissions for the packet is not reached. On the other hand, when the maximum number of transmissions for the packet is reached, the packet will be removed from the buffer without the need of ACK or NACK. Therefore, outage occurs at the maximum round of transmission, i.e.,  $M^{\text{th}}$  transmission, if the packet is discarded from the buffer while the receiver does not correctly receive this packet.

We can model the buffer activity at the end of each frame as a discrete-time Markov process [21]. Fig. 2 depicts the state transition model. State 0 denotes that the packet is removed from the buffer, and state  $m$  represents the number of retransmissions for the packet, where no packet is removed from the buffer. Define  $p_m$  as the decoding failure probability at the  $m^{\text{th}}$  retransmission such that the system enters State  $m+1$  with probability  $p_m$ , while the system enters state 0 with probability  $1-p_m$ . On the other hand, regardless of the decoding result at the end of  $(M-1)^{\text{th}}$  retransmission, the system goes to State 0 with probability  $1 = p_{M-1} + (1-p_{M-1})$  since the maximum number of transmissions is reached and the packet is removed immediately from the

buffer. Then, the state transition matrix is given by

$$\mathbf{P} = \begin{bmatrix} 1 - p_0 & 1 - p_1 & \cdots & 1 - p_{M-2} & 1 \\ p_0 & 0 & \cdots & 0 & 0 \\ 0 & p_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{M-2} & 0 \end{bmatrix} \quad (3)$$

where  $P_{ij}$  denotes the probability of state transition from State  $j$  to State  $i$ .

In the HARQ-IR protocol, the transmitter encodes the packet according to a codebook of length  $MTB$ , and the codewords are divided into subblocks of the same length with  $TB$  symbols. During each frame, only one subblock is sent to the receiver, and the receiver decodes the message using the current subblock combined with the previously received subblocks of the packet. Then, we know that the receiver can successfully decode the packet after the  $m^{\text{th}}$  ( $0 \leq m \leq M - 1$ ) retransmission only if the following condition is satisfied [2]

$$L \leq \sum_{i=0}^m TB \log_2 (1 + \text{SNR} z_i). \quad (4)$$

We can express the state transition probabilities as  $p_0 = \Pr \left\{ z < \frac{2^{L/TB} - 1}{\text{SNR}} \right\}$ , and for  $m = 1, \dots, M - 1$ ,

$$p_m = \frac{\Pr \{ L > \sum_{i=0}^m TB \log_2 (1 + \text{SNR} z_i) \}}{\Pr \{ L > \sum_{i=0}^{m-1} TB \log_2 (1 + \text{SNR} z_i) \}}. \quad (5)$$

Define the outage probability after  $m$ -th transmission rounds as

$$P_{\text{out},m} = \Pr \left\{ \sum_{i=0}^{m-1} TB \log_2 (1 + \text{SNR} z_i) < L \right\}. \quad (6)$$

In the absence of delay constraints, the throughput of truncated HARQ, i.e., *goodput*, is known to be [2]

$$R_{\text{HARQ}} = \frac{L}{TB} \frac{(1 - P_{\text{out},M})}{\sum_{m=0}^{M-1} P_{\text{out},m}}, \text{ bps/Hz}, \quad (7)$$

where  $P_{\text{out},0} = 1$ .

### III. EFFECTIVE CAPACITY ANALYSIS IN ONE-HOP LINKS

In this section, we first review the preliminaries on the statistical delay constraints, and then obtain the outage effective capacity for a communication link with the parameters discussed above.

#### A. Statistical Delay Constraints for One-Hop Links

Suppose that the queue is stable and that both the arrival process  $a[n]$  and service process  $c[n]$  satisfy the Gärtner-Ellis limit, i.e., for all  $\theta \geq 0$ , there exists a differentiable logarithmic moment generating function (LMGF)  $\Lambda_A(\theta)$  such that<sup>1</sup>  $\lim_{n \rightarrow \infty} \frac{\log \mathbb{E}\{e^{\theta \sum_{i=1}^n a[i]}\}}{n} = \Lambda_A(\theta)$ , and a differentiable LMGF  $\Lambda_C(\theta)$  such that  $\lim_{n \rightarrow \infty} \frac{\log \mathbb{E}\{e^{\theta \sum_{i=1}^n c[i]}\}}{n} = \Lambda_C(\theta)$ . If there exists a unique  $\theta^* > 0$  such that

$$\Lambda_A(\theta^*) + \Lambda_C(-\theta^*) = 0, \quad (8)$$

then [26]

$$\lim_{Q_{\max} \rightarrow \infty} \frac{\log \Pr\{Q > Q_{\max}\}}{Q_{\max}} = -\theta^*. \quad (9)$$

where  $Q$  is the stationary queue length.

For large  $Q_{\max}$ , we have the approximation for the buffer violation probability:  $\Pr\{Q > Q_{\max}\} \approx e^{-\theta^* Q_{\max}}$ . Hence, while larger  $\theta$  corresponds to stricter queueing constraints, smaller  $\theta$  implies looser queueing constraints. Then, equivalently, we have the queueing delay violation probability as  $\Pr\{D > D_{\max}\} \approx e^{-J(\theta) D_{\max}}$ , where

$$J(\theta) = -\Lambda_C(-\theta)$$

is the statistical delay exponent associated with the queue, with  $\Lambda_C(\theta)$  the LMGF of the service rate. Then, the maximum constant arrival rate to the queue for given  $\theta > 0$  is expressed as

$$R_E(\theta) = -\frac{\Lambda_C(-\theta)}{\theta T B}, \text{ bps/Hz.} \quad (10)$$

<sup>1</sup>Throughout the text, logarithm expressed without a base, i.e.,  $\log(\cdot)$ , refers to the natural logarithm  $\log_e(\cdot)$ .



### B. Outage Effective Capacity

While the authors in [21] considered the departure processes of the source queue, we focus on the *goodput* departure processes that can be correctly received at the receiver similar to [22]. According to (8), we define the outage effective capacity as the maximum constant arrival rate to the source that can be supported by the *goodput* processes. Then, we can obtain the following result.

*Theorem 1:* For the fixed rate transmissions with HARQ protocols, given QoS exponent  $\theta > 0$ , SNR  $> 0$ , and maximum number of transmissions  $M$ , the outage effective capacity is given by

$$R_{\text{out}}(\text{SNR}, \theta) = \frac{1}{TB} \max_{L \geq 0} \left\{ -\frac{\Lambda(-\theta)}{\theta} \right\} \quad (11)$$

$$= \max_{L \geq 0} \left\{ -\frac{1}{\theta TB} \log(p_0 (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L}) y^*) \right\} \quad (12)$$

$$= -\frac{1}{\theta TB} \log(p_{0,\text{opt}} (P_{\text{out},\text{opt}} + (1 - P_{\text{out},\text{opt}})e^{-\theta L_{\text{opt}}}) y_{\text{opt}}^*) \quad (13)$$

where  $L_{\text{opt}}$  is the optimal finite fixed transmission rate that solves (12),  $y^*$  is the only unique real positive root of  $f(y) = 0$  with

$$f(y) = y^M - \frac{1 - p_0}{p_0} y^{M-1} - \sum_{m=1}^{M-2} \frac{(1 - p_m) p_{m-1} \cdots p_1}{p_0^m (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})^m} y^{M-1-m} - \frac{p_{M-2} \cdots p_1}{p_0^{M-1} (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})^{M-1}} \quad (14)$$

for given  $L$ , and  $y_{\text{opt}}^*$  is the only unique real positive root of  $f(y) = 0$  with  $L = L_{\text{opt}}$ . The outage probability associated can be expressed as

$$P_{\text{out},\text{opt}} = \prod_{m=0}^{M-1} p_{m,\text{opt}} = \Pr \left\{ \sum_{m=0}^{M-1} TB \log_2(1 + \text{SNR} z_m) < L_{\text{opt}} \right\}$$

where  $p_{m,\text{opt}}$  denotes the state transition probability obtained with  $L_{\text{opt}}$ .

*Proof:* See Appendix A. □

*Remark 1:* Above, we did not specify how  $L_{\text{opt}}$  is obtained. Since there is no closed form expression for  $y^*$  which depends on  $L$  nonlinearly, we can only solve (12) numerically. For instance, in the following numerical results, we employ branch-and-bound method to find  $L_{\text{opt}}$ . In general,  $L_{\text{opt}}$  depends on  $\theta$ , SNR, and  $M$ . Note that the rate expression in (12) is applicable for all  $\theta > 0$ , in stark difference from the results in [20], where the closed-form expression of

effective capacity is obtained for small  $\theta$ . Also, we characterized the outage probability that was not treated in [20]. Note also that the rate expression in (13) is different from the results in [21], where packet drop is not considered, and the results in [22], where the results are in matrix form based on a random walk model and recurrence relation formulation.

In the absence of statistical QoS constraints, we have the following result.

*Proposition 1:* As  $\theta \rightarrow 0$ , the outage effective capacity with HARQ protocols is given by

$$\lim_{\theta \rightarrow 0} R_{\text{out}}(\text{SNR}, \theta) = \frac{L_{\text{opt}}}{TB} \frac{(1 - P_{\text{out,opt},M})}{\sum_{m=0}^{M-1} P_{\text{out,opt},m}}, \text{ bps/Hz.} \quad (15)$$

*Proof:* See Appendix B. □

*Remark 2:* Note that the outage effective capacity approaches to the maximum goodput of the HARQ protocols, i.e.,  $\max_{L \geq 0} R_{\text{HARQ}}$ , as the statistical QoS constraints vanish.

*Remark 3:* The results in Theorem 1 is generic, and can be applied to other HARQ protocols as well, e.g., HARQ-T1, where the transmitter sends the same packet in each frame during retransmissions and the receiver decodes the packet successfully if the instantaneous channel rate is greater than the transmission rate, and HARQ-chase combining (HARQ-CC), where the receiver can make use of the received signals in the previous frames through maximum ratio combining. Note that it has been verified that HARQ-IR performs better than the other schemes under statistical delay constraints [20]. As  $M \rightarrow \infty$ , the outage probability vanishes and the outage effective capacity is exactly the constant arrival rate supported by the departure processes. Here, we give an example of HARQ-T1 when  $M \rightarrow \infty$ .

*Proposition 2:* For the fixed rate transmissions with HARQ-T1 protocol, the outage effective capacity for a given QoS exponent  $\theta > 0$  and  $\text{SNR} > 0$  approaches to the following value as  $M \rightarrow \infty$ :

$$\lim_{M \rightarrow \infty} R_{\text{out}}(\text{SNR}, \theta) = \max_{L \geq 0} \left\{ -\frac{1}{\theta TB} \log \left( 1 - \Pr \left\{ z > \frac{2^{L/TB} - 1}{\text{SNR}} \right\} (1 - e^{-\theta L}) \right) \right\}. \quad (16)$$

*Proof:* See Appendix C. □

Obviously, (16) coincides with the result in [17, (11)].

#### IV. OUTAGE EFFECTIVE CAPACITY IN BUFFER-AIDED DIAMOND RELAY SYSTEMS

In this section, we first briefly discuss the statistical delay constraints for two-hop links, and then define and characterize the outage effective capacity for the buffer-aided diamond relay

systems under consideration.

#### A. Statistical Delay Constraints for Two-Hop Links

In this work, we seek to identify the maximum constant arrival rate to the source that can be supported by the *goodput* processes successively received at the destination of the diamond relay system using HARQ-IR while satisfying the statistical delay constraints. Therefore, we need to guarantee that the data transmission of all information flows should satisfy the statistical delay constraints. Since there is *no link* between the relays, we have at most two concatenated queues for the information flow. Consider two concatenated queues with statistical queueing constraints specified by  $\theta_1$  and  $\theta_2$ , for queue 1 and queue 2, respectively. Given the queueing constraints specified by  $\theta_1$  and  $\theta_2$  with (9) satisfied for each queue, we define

$$J_1(\theta_1) = -\Lambda_{C,1}(-\theta_1), \text{ and } J_2(\theta_2) = -\Lambda_{C,2}(-\theta_2), \quad (17)$$

where  $\Lambda_{C,1}(\theta_1)$  and  $\Lambda_{C,2}(\theta_1)$  are the LMGF functions of the service rate of queue 1, 2, respectively. For data going through both queues, the end-to-end queueing delay violation probability can be characterized as

$$\begin{aligned} \Pr\{D_1 + D_2 > D_{\max}\} &\doteq 1 - \int_0^{D_{\max}} \int_0^{D_{\max}-D_1} p_D(D_1)p_D(D_2)dD_2dD_1 \\ &= \begin{cases} \frac{J_1(\theta_1)e^{-J_2(\theta_2)D_{\max}} - J_2(\theta_2)e^{-J_1(\theta_1)D_{\max}}}{J_1(\theta_1) - J_2(\theta_2)}, & J_1(\theta_1) \neq J_2(\theta_2) \\ (1 + J_1(\theta_1)D_{\max})e^{-J_1(\theta_1)D_{\max}}, & J_1(\theta_1) = J_2(\theta_2). \end{cases} \end{aligned} \quad (18)$$

Thereby, we need to guarantee that

$$\Pr\{D_1 + D_2 > D_{\max}\} \leq \varepsilon. \quad (19)$$

In this way, we can guarantee that the data transmissions through the relays, i.e., information flows over two queues at the source and the relays, satisfy the statistical delay constraints. Then, the delay constraints of the whole system can be satisfied. Note that  $(\varepsilon, D_{\max})$  characterizes the statistical delay constraints with maximum delay violation probability  $\varepsilon$  and maximum delay  $D_{\max}$ . To facilitate the following analysis, we need the following tradeoff between the delay exponents of any concatenated two queues, i.e.,  $J_1(\theta_1)$  and  $J_2(\theta_2)$ .

*Lemma 1 ([23]):* Consider the following function

$$\vartheta(J_1(\theta_1), J_2(\theta_2)) = \frac{J_2(\theta_2)e^{-J_1(\theta_1)D_{\max}} - J_1(\theta_1)e^{-J_2(\theta_2)D_{\max}}}{J_2(\theta_2) - J_1(\theta_1)} = e^{-J_0 D_{\max}} = \varepsilon, \text{ for } 0 \leq \varepsilon \leq 1, \quad (20)$$

where  $J_0 = -\frac{\log(\varepsilon)}{D_{\max}}$  is defined as the statistical delay exponent associated with  $(\varepsilon, D_{\max})$ . Denoting  $J_2(\theta_2) = \Phi(J_1(\theta_1))$  as a function of  $J_1(\theta_1)$ , we have

a)  $\Phi$  is continuous. For  $J_1(\theta_1) = J_{th}(\varepsilon)$ , we have

$$\Phi(J_1(\theta_1)) = J_{th}(\varepsilon), \quad (21)$$

where

$$J_{th}(\varepsilon) = -\frac{1}{D_{\max}} \left( 1 + \mathcal{W}_{-1} \left( -\frac{\varepsilon}{e} \right) \right), \quad (22)$$

where  $\mathcal{W}_{-1}(\cdot)$  is the Lambert W function, which is the inverse function of  $y = xe^x$  in the range  $(-\infty, -1]$ .

b)  $\Phi$  is strictly decreasing in  $J_1(\theta_1)$ .

c)  $\Phi$  is convex in  $J_1(\theta_1)$ .

d)  $J_1(\theta_1) \in [J_0, \infty)$ , and  $J_2(\theta_2) = \Phi(J_1(\theta_1)) \in [J_0, \infty)$ .

### B. Effective Capacity Analysis of Diamond-Relay Systems

If we define  $\theta_1$ ,  $\theta_{r_1}$  and  $\theta_{r_2}$  as the statistical queueing constraints at the source and the relays, respectively. For different information flows over the relays, we will have different two-hop channels with queueing constraints  $(\theta_1, \theta_{r_1})$  and  $(\theta_1, \theta_{r_2})$ , respectively. Assume that the equivalent constant arrival rate at the source is  $R \geq 0$ . Consider any realization  $(\theta_1, \theta_2)$  of any two concatenated queues. Denote  $\Omega$  as the set of pairs  $(\theta_1, \theta_2)$  such that (19) can be satisfied. To satisfy the queueing constraint at queue 1, i.e., queue at the source, we should have  $\tilde{\theta} \geq \theta_1$ , where  $\tilde{\theta}$  is the solution to

$$R = -\frac{\Lambda_{C,1}(-\tilde{\theta})}{\tilde{\theta}}, \quad (23)$$

and  $\Lambda_{C,1}(\theta)$  is the LMGF of the *goodput* service for queue 1, i.e., service processes successively received at queue 2 (any relay).

Also, in order to satisfy the queueing constraint of queue 2, we must have  $\hat{\theta} \geq \theta_2$ , where  $\hat{\theta}$  is the solution to

$$\Lambda_{A,2}(\theta) + \Lambda_{C,2}(-\theta) = 0. \quad (24)$$

where  $\Lambda_{A,2}(\theta)$  is the LMGF of the *goodput* arrivals at queue 2,  $\Lambda_{C,2}(\theta)$  is the LMGF of the *goodput* service of queue 2, i.e., service processes successively received at destination. Note that we need to consider the queues at the relays together with the queue at the source.

Denote  $\Omega$  as the set of pairs  $(\theta_1, \theta_2)$  of two concatenated buffers such that (19) can be satisfied. Now, *outage effective capacity* of the buffer-aided diamond relay system under statistical delay constraints  $(\varepsilon, D_{\max})$  can be formulated as follows.

*Definition 1:* The outage effective capacity of the buffer-aided diamond relay system with statistical delay constraints specified by  $(\varepsilon, D_{\max})$  is given by

$$R(\varepsilon, D_{\max}) = \sup_{(\theta_1, \theta_{r_1}) \in \Omega, (\theta_1, \theta_{r_2}) \in \Omega} R. \quad (25)$$

Hence, outage effective capacity is now the maximum constant arrival rate that can be supported by the *goodput* processes successfully received at the destination of the diamond relay system under statistical delay constraints.

### C. Outage Effective Capacity of Diamon-Relay Links with HARQ-IR

In this part, we study the performance of HARQ-IR in the buffer-aided diamond-relay channels. We assume that common messages are sent to the relays and the relays cooperate in the information delivery to the destination such that the queue dynamics at the relays are the same. We consider the *end-to-end* delay constraints, and identify the maximum constant arrival goodput to the source and the end-to-end outage probability while satisfying the statistical delay constraints.

1) *Decode-and-Forward (DF):* As a comparison, we consider the decode-and-forward (DF) scheme [12], in which case the CSI is also available at the transmitter for each link and each relay must successfully decode the common message transmitted by the source node, and later the relays can cooperatively beamform their transmissions to the destination. We assume that the transmission power levels at the source and relays are fixed and hence no power control

is employed (i.e., nodes are subject to short-term power constraints). We further assume that the channel capacity for each link can be achieved, i.e., the service processes are equal to the instantaneous Shannon capacities of the links such that there is no decoding error. Then, the service rate leaving the queue at the source is given by

$$C_s = TB \log_2(1 + \text{SNR}_s \min\{z_{sr_1}, z_{sr_2}\}). \quad (26)$$

Also, the rates leaving the queues at the relays are the same, and are given by

$$C_{r_1} = C_{r_2} = TB \log_2 \left( 1 + (\sqrt{\text{SNR}_{r_1} z_{r_1d}} + \sqrt{\text{SNR}_{r_2} z_{r_2d}})^2 \right). \quad (27)$$

Above, the rates are given in terms of bits/block. Note that the arrival rates and departure rates of the queues at the relays are always the same, and hence the queueing activities have the same pattern. Therefore, the system simplifies to the two-hop channel. Then, we can obtain the effective capacity similar to the discussions in [23]. In this scheme, the end-to-end outage probability is zero, i.e., all departure processes can be successfully received at the destination.

2) *HARQ-IR*: We assume perfect CSI is available only at the receiver for each link, in which case HARQ-IR is incorporated for the transmissions. Similar to the discussion in Section II-B, we first assume that a packet of  $L$  bits is intended to be transmitted in each frame for the each hop and obtain the outage effective capacity associated with  $L$ . Then, we optimize over  $L \geq 0$  to find the optimal  $L_{\text{opt}}$  that leads to the maximum outage effective capacity.

The operations of HARQ-IR can be described as follows:

- a) In the first hop, the source tries to send the same information to the relays. Note that only after reception of ACKs from all relays, the packet can be removed from the buffer, and the source attempts to send  $L$  bits in the next frame. Again, we model the source buffer activity at the end of each frame as a discrete-time Markov process. Define  $p_{s,m}$  as the decoding failure probability at the  $m^{\text{th}}$  retransmission such that the system enters State  $m+1$  with probability  $p_{s,m}$ , while the system enters state 0 with probability  $1 - p_{s,m}$ . On the other hand, regardless of the decoding result at the end of  $(M-1)^{\text{th}}$  retransmission, the system goes to State 0 with probability  $1 = p_{s,M-1} + (1 - p_{s,M-1})$  since the maximum number of transmissions is reached and the packet is removed immediately from the source buffer. The state transition matrix  $\mathbf{P}_s$  can be expressed similar to (3) with values  $p_{s,m}$  instead. For each relay, we know that

the relay can successfully decode the packet after the  $m^{\text{th}}$  ( $0 \leq m \leq M - 1$ ) retransmission only if the following condition is satisfied

$$L \leq \sum_{i=0}^m TB \log_2 (1 + \text{SNR} z_{sr_j, i}), j = 1, 2. \quad (28)$$

Therefore, we can express the state transition probabilities as

$$\begin{aligned} p_{s,0} &= \Pr \left\{ \left\{ z_{sr_1} < \frac{2^{L/TB} - 1}{\text{SNR}} \right\} \cup \left\{ z_{sr_2} < \frac{2^{L/TB} - 1}{\text{SNR}} \right\} \right\} \\ &= 1 - \Pr \left\{ z_{sr_1} \geq \frac{2^{L/TB} - 1}{\text{SNR}} \right\} \Pr \left\{ z_{sr_2} \geq \frac{2^{L/TB} - 1}{\text{SNR}} \right\} \end{aligned} \quad (29)$$

and for  $m = 1, \dots, M - 1$ , we have

$$p_{s,m} = \frac{\Pr \{ \{ L > \sum_{i=0}^m TB \log_2 (1 + \text{SNR} z_{sr_1, i}) \} \cup \{ L > \sum_{i=0}^m TB \log_2 (1 + \text{SNR} z_{sr_2, i}) \} \}}{\Pr \{ \{ L > \sum_{i=0}^{m-1} TB \log_2 (1 + \text{SNR} z_{sr_1, i}) \} \cup \{ L > \sum_{i=0}^{m-1} TB \log_2 (1 + \text{SNR} z_{sr_2, i}) \} \}} \quad (30)$$

$$= \frac{1 - \Pr \{ L \leq \sum_{i=0}^m TB \log_2 (1 + \text{SNR} z_{sr_1, i}) \} \Pr \{ L \leq \sum_{i=0}^m TB \log_2 (1 + \text{SNR} z_{sr_2, i}) \}}{1 - \Pr \{ L \leq \sum_{i=0}^{m-1} TB \log_2 (1 + \text{SNR} z_{sr_1, i}) \} \Pr \{ L \leq \sum_{i=0}^{m-1} TB \log_2 (1 + \text{SNR} z_{sr_2, i}) \}}. \quad (31)$$

- b) In the second hop, the relays attempt to send the same message to the destination. Following the idea of treating the relays as distributed antennas, we can adopt the Alamouti scheme to improve the achievable rate. Specifically, we divide the frame into two slots of equal length  $TB/2$ . In one slot, the relay  $\mathbf{R}_1$  sends message  $x_1$ , and the relay  $\mathbf{R}_2$  sends message  $x_2$ . In the other slot, the relay  $\mathbf{R}_1$  sends message  $x_2^*$ , and the relay  $\mathbf{R}_2$  sends message  $-x_1^*$ . Then, the achievable rate for the second hop in each frame can be expressed as

$$R = TB \log_2 (1 + \text{SNR}_{r_1} z_{r_1d} + \text{SNR}_{r_2} z_{r_2d}), \text{ bits/block}. \quad (32)$$

Note that the arrival rates and departure rates of the queues at the relays are always the same, and hence the queueing activities have the same pattern. Now, for the Markov process associated with the buffer activities at the relays, we have the state transition matrix  $\mathbf{P}_r$  with state transition probabilities as  $p_{r,0} = \Pr \{ \text{SNR}_{r_1} z_{r_1d} + \text{SNR}_{r_2} z_{r_2d} < 2^{L/TB} - 1 \}$ , and for

$$m = 1, \dots, M - 1,$$

$$p_{r,m} = \frac{\Pr \{L > \sum_{i=0}^m TB \log_2 (1 + \text{SNR}_{r_1} z_{r_1 d,i} + \text{SNR}_{r_2} z_{r_2 d,i})\}}{\Pr \{L > \sum_{i=0}^{m-1} TB \log_2 (1 + \text{SNR}_{r_1} z_{r_1 d,i} + \text{SNR}_{r_2} z_{r_2 d,i})\}}.$$

We can obtain the statistical delay exponent for each hop as

$$J_1(\theta_1) = -\Lambda_{C,1}(-\theta_1) = -\log sp \{ \mathbf{P}_s \phi_s(-\theta_1) \} \quad (33)$$

$$J_2(\theta_2) = -\Lambda_{C,2}(-\theta_2) = -\log sp \{ \mathbf{P}_r \phi_r(-\theta_2) \} \quad (34)$$

where  $\phi_s(\theta_1) = \text{diag}(P_{\text{out},s} + (1 - P_{\text{out},s})e^{\theta_1}, \underbrace{1, \dots, 1}_{M-1})$  and  $\phi_r(\theta_2) = \text{diag}(P_{\text{out},r} + (1 - P_{\text{out},r})e^{\theta_2}, \underbrace{1, \dots, 1}_{M-1})$  are diagonal matrices with each component given by the moment generating functions of the *goodput* processes in  $M$  states of the Markov processes  $\mathbf{P}_s$  and  $\mathbf{P}_r$ , where  $P_{\text{out},k} = \prod_{m=0}^{M-1} p_{k,m}$ ,  $k = s, r$  denotes the outage probability of the first and second hop, respectively.

Given  $L > 0$ , we denote  $J_{1,\max} = \lim_{\theta_1 \rightarrow \infty} J_1(\theta_1)$  and  $J_{2,\max} = \lim_{\theta_2 \rightarrow \infty} J_2(\theta_2)$  as the maximum delay exponent of the first and second hop, which is obtained as the statistical queueing constraints approach infinity. We can show the following results.

*Proposition 3:* With the HARQ-IR protocol,  $J_{1,\max}$  and  $J_{2,\max}$  are finite if  $p_{k,0} \neq 0, k = s, r$ .

*Proof:* First, it can be easily verified that  $P_{\text{out},k} \neq 0, k = s, r$  if  $p_{k,0} \neq 0, k = s, r$  since  $p_{k,m} \neq 0, k = s, r, m = 1, \dots, M - 1$ . As  $\theta \rightarrow \infty$ , we can see from (14) that  $y^*$  will be the solution to the following equation

$$\lim_{\theta \rightarrow \infty} f(y) = y^M - \frac{1 - p_0}{p_0} y^{M-1} - \sum_{m=1}^{M-2} \frac{(1 - p_m) p_{m-1} \cdots p_1}{p_0^m P_{\text{out}}^m} y^{M-1-m} - \frac{p_{M-2} \cdots p_1}{p_0^{M-1} P_{\text{out}}} = 0. \quad (35)$$

Obviously,  $y^*$  approaches to some finite value. Hence,  $\lim_{\theta \rightarrow \infty} J(\theta) = \lim_{\theta \rightarrow \infty} -\Lambda(-\theta) = -\log(p_0 P_{\text{out}} y^*)$  is finite, which implies that  $J_{1,\max}$  and  $J_{2,\max}$  are finite.  $\square$

*Remark 4:* Note that  $p_0 \neq 0$  means that the possibility of failure to decode the packet in the first transmission is not zero. For the fading distributions such as Rayleigh and Nakagami-m, we can see that  $p_0 = \Pr \left\{ z < \frac{2^{L/TB} - 1}{\text{SNR}} \right\}$  is greater than zero for all  $L > 0$ .

Define

$$\Omega_\varepsilon = \{(\theta_1, \theta_2) : J_1(\theta_1) \text{ and } J_2(\theta_2) \text{ are solutions to (19) w/ equality}\}.$$

With the above characterizations, we can obtain the following results.



*Theorem 2:* Given  $L > 0$ , the outage effective capacity of the buffer-aided diamond relay systems with HARQ-IR strategy subject to statistical delay constraints specified by  $(\varepsilon, D_{\max})$  is given by the following:

**Case I:** If  $\vartheta(J_{1,\max}, J_{2,\max}) > \varepsilon$ ,

$$R_{\text{HARQ-IR}}(\varepsilon, D_{\max}, L) = 0, \quad (36)$$

**Case II:** Otherwise,

**Case II.a:** If  $J_{1,\max} < J_{th}(\varepsilon)$ ,

$$R_{\text{HARQ-IR}}(\varepsilon, D_{\max}, L) = \frac{J_1(\overset{\circ}{\theta}_1)}{\overset{\circ}{\theta}_1}, \quad (37)$$

where  $\overset{\circ}{\theta}_1$  is the smallest value of  $\theta_1$  with  $(\theta_1, \theta_2) \in \Omega_\varepsilon$  satisfying

$$J_1(\theta_1) = J_2(\theta_2) + J_1(\theta_1 - \theta_2). \quad (38)$$

**Case II.b:** If  $J_{2,\max} < J_{th}(\varepsilon)$ ,

$$R_{\text{HARQ-IR}}(\varepsilon, D_{\max}, L) = \frac{J_2(\check{\theta}_2)}{\check{\theta}_2} \quad (39)$$

where  $(\check{\theta}_1, \check{\theta}_2)$  is the unique solution to

$$\frac{J_1(\theta_1)}{\theta_1} = \frac{J_2(\theta_2)}{\theta_2}, \quad (40)$$

with  $(\theta_1, \theta_2) \in \Omega_\varepsilon$ .

**Case II.c:** If  $J_{1,\max} \geq J_{th}(\varepsilon)$  and  $J_{2,\max} \geq J_{th}(\varepsilon)$ ,

a) If  $\theta_{1,th} = \theta_{2,th}$ ,

$$R_{\text{HARQ-IR}}(\varepsilon, D_{\max}, L) = \frac{J_{th}(\varepsilon)}{\theta_{1,th}}, \quad (41)$$

where  $(\theta_{1,th}, \theta_{2,th})$  is the unique solution pair to  $J_1(\theta_1) = J_{th}(\varepsilon)$ , and  $J_2(\theta_2) = J_{th}(\varepsilon)$ .

b) If  $\theta_{1,th} > \theta_{2,th}$ ,

$$R_{\text{HARQ-IR}}(\varepsilon, D_{\max}, L) = \frac{J_1(\overset{\circ}{\theta}_1)}{\overset{\circ}{\theta}_1} \quad (42)$$

where  $\overset{\circ}{\theta}_1$  is the smallest value of  $\theta_1$  with  $(\theta_1, \theta_2) \in \Omega_\varepsilon$  satisfying

$$J_1(\theta_1) = J_2(\theta_2) + J_1(\theta_1 - \theta_2) \quad (43)$$

c) If  $\theta_{1,th} < \theta_{2,th}$ ,

$$R_{HARQ-IR}(\varepsilon, D_{\max}, L) = \frac{J_2(\check{\theta}_2)}{\check{\theta}_2} \quad (44)$$

where  $(\check{\theta}_1, \check{\theta}_2)$  is the unique solution to

$$\frac{J_1(\theta_1)}{\theta_1} = \frac{J_2(\theta_2)}{\theta_2}, \quad (45)$$

with  $(\theta_1, \theta_2) \in \Omega_\varepsilon$ .

The associated end-to-end outage probability is given by

$$P_{\text{out}} = 1 - (1 - P_{\text{out},s})(1 - P_{\text{out},r}). \quad (46)$$

*Proof:* See Appendix D. □

*Remark 5:* Note that due to the outage events, it is possible that certain delay constraints may not be satisfied, e.g., **Case I**.

*Proposition 4:* The outage effective capacity of the buffer-aided diamond relay systems with HARQ-IR strategy subject to statistical delay constraints specified by  $(\varepsilon, D_{\max})$  can be expressed as

$$R_{HARQ-IR}(\varepsilon, D_{\max}) = \max_{L \geq 0} R_{HARQ-IR}(\varepsilon, D_{\max}, L) = R_{HARQ-IR}(\varepsilon, D_{\max}, L_{\text{opt}}). \quad (47)$$

The associated optimal end-to-end outage probability for  $L_{\text{opt}}$  is given by

$$P_{\text{out,opt}} = 1 - (1 - P_{\text{out,opt},s})(1 - P_{\text{out,opt},r}). \quad (48)$$

*Remark 6:* Following the similar reasoning in Appendix A, we can show that the outage effective capacity approaches to 0 when  $L \rightarrow 0$  or  $L \rightarrow \infty$ . So  $L_{\text{opt}}$  is finite, and the approach in Remark 1 can be used here to derive  $L_{\text{opt}}$ .

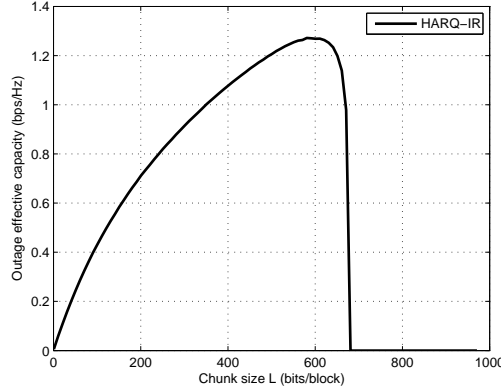


Fig. 3. The outage effective capacity as  $L$  varies.  $M = 4$ .

## V. NUMERICAL RESULTS

In the numerical results, we assume the fading distributions of all links follow independent Rayleigh fading with means  $\mathbb{E}\{z_{sr_j}\} = \mathbb{E}\{z_{r_jd}\} = 16$ ,  $j = 1, 2$ ,  $\text{SNR} = 0$  dB,  $T = 1$  ms,  $B = 180$  kHz, and  $D_{\max} = 1$  s. Now,  $J_{1,\max}$  and  $J_{2,\max}$  are finite from Remark 4.

In Fig. 3, we plot the outage effective capacity as a function of  $L$ . In this figure, we assume  $\text{SNR}_r = 5$  dB,  $M = 4$ , and  $\varepsilon = 0.05$ . We can see that the outage effective capacity is maximized at a finite value  $L_{\text{opt}}$ . Also, we can find that when  $L$  is larger than certain value, the outage effective capacity vanishes immediately. This is due to the fact that when  $L$  is large enough, the outage probability of each hop can be so large that the end-to-end delay constraints cannot be satisfied, i.e., **Case I** of Theorem 2.

In Fig. 4, we plot the outage effective capacity as a function of  $\text{SNR}_r$ . From the figure, it is interesting that HARQ-IR based transmission scheme can achieve larger effective capacity compared with DF protocol at relatively small  $\text{SNR}_r$  levels at the relays, albeit at the expense of outage. This is generally due to the fact that at smaller  $\text{SNR}_r$  values, the effective capacity is maximized at larger  $J_1(\theta_1)$ , or larger  $\theta_1$  equivalently. In this case, the system enjoys the benefit of average over different channel realizations provided by HARQ-IR, which can lead to larger effective capacity. On the other hand, when  $\text{SNR}_r$  becomes large, we can see from (27) that the service rate of the second hop of DF protocol increases significantly compared with the one achieved with the HARQ-IR protocol in (32), which will result in much looser delay constraints

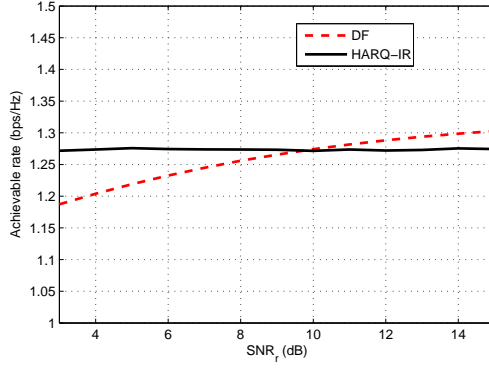


Fig. 4. Effective capacity as a function of  $\text{SNR}_r$ .  $\varepsilon = 0.05$ .

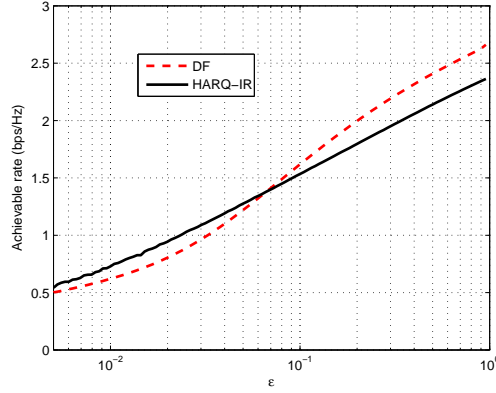


Fig. 5. Effective capacity as a function of  $\varepsilon$ .  $\text{SNR}_r = 5$  dB.

$J_1(\theta_1)$  at the source, i.e., smaller  $\theta_1$ , and hence the effective capacity of DF protocol is larger.

In Fig. 5, we plot the outage effective capacity as  $\varepsilon$  varies. We assume  $\text{SNR}_r = 5$  dB. We can find that the HARQ-IR based scheme achieves superior performance than DF protocol when  $\varepsilon$  is relative small, i.e., stringent *end-to-end* delay constraints. The reasoning behind is similar to previous finding. That is, at relative large  $\theta_1$ , the benefit provided by averaging over different channel realizations with HARQ-IR is more prominent. In Fig. 6, we plot the associated outage probability as  $\varepsilon$  varies. We can find that as the delay constraints become more stringent, i.e.,  $\varepsilon$  decreases, the outage probability decreases. This is obvious since smaller outage probability implies less retransmissions to avoid build-up in the buffers. It is interesting that the optimal outage probability appears to be linear in the delay violation probability.

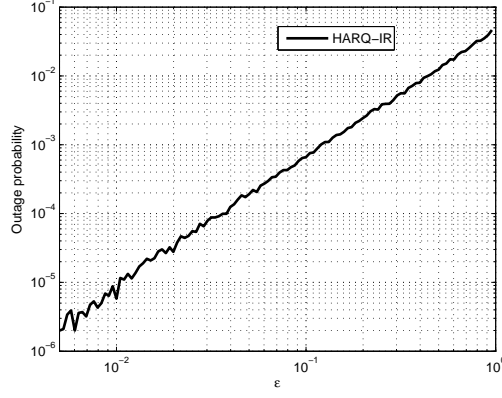


Fig. 6. Outage probability as a function of  $\varepsilon$ .  $\text{SNR}_r = 5$  dB.

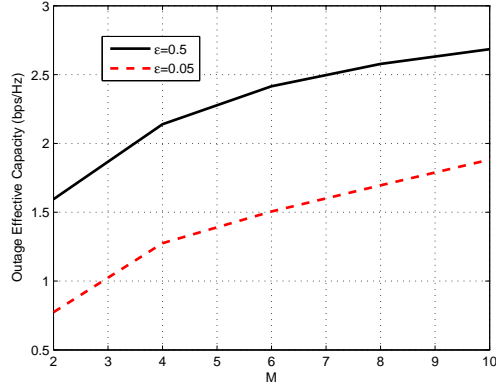


Fig. 7. The outage effective capacity vs.  $M$ .

In Fig. 7, we plot the outage effective capacity as a function of  $M$ . We assume  $\text{SNR}_r = 5$  dB and  $\varepsilon = \{0.5, 0.05\}$ . From the figure, we can see that the outage effective capacity of the buffer-aided diamond relay systems using HARQ-IR is increasing in  $M$ , similar to the findings in [20] for one-hop links.

## VI. CONCLUSION

In this paper, we have investigated the buffer-aided diamond relay systems with truncated HARQ-IR protocol under delay constraints. We have assumed that there is only perfect CSI at the receiver side for each link, and the transmitters send the information at a fixed rate. We

have introduced the notion of *outage effective capacity*, which identifies the maximum constant rate to the transmitter that can be supported by the goodput processes correctly received at the destination. We have characterized the outage effective capacity and the associated outage probability in buffer-aided diamond relay systems with HARQ-IR. Through numerical results, we have found that HARQ-IR achieves better performance than the DF protocol with perfect CSI at the transmitters as well when the SNR at the relays are relatively small or when the delay constraints are stringent. It is interesting that the optimal end-to-end outage probability appears to be linear in the delay violation probability.

## APPENDIX

### A. Proof of Theorem 1

Since we consider the *goodput* of the departures that can be successively received at the receiver, for the state transition model in Fig. 2, outage occurs when the departure in State  $M - 1$  cannot be correctly received at the receiver. Then, such departure processes contribute nothing to the *goodput*. Note that outage occurs only at the State  $M - 1$ , i.e., decoding failure after the  $M^{\text{th}}$  transmission of the packet. We know that with fixed transmission rate  $L$ , the outage probability is given by

$$\begin{aligned}
P_{\text{out}} &= \Pr\{\text{decoding failure after the } M^{\text{th}} \text{ transmission of the packet}\} \\
&= \Pr\{\text{decoding failure after the } 1^{\text{st}} \text{ transmission of the packet}\} \\
&\quad \times \Pr\{\text{decoding failure after the } 2^{\text{nd}} \text{ transmission of the packet} \\
&\quad \quad |\text{decoding failure after the } 1^{\text{st}} \text{ transmission of the packet}\} \\
&\quad \times \cdots \times \Pr\{\text{decoding failure after the } M^{\text{th}} \text{ transmission of the packet} \\
&\quad \quad |\text{decoding failure after the } (M - 1)^{\text{th}} \text{ transmission of the packet}\} \\
&= p_0 \times p_1 \times \cdots \times p_{M-1} = \prod_{m=0}^{M-1} p_m \tag{49}
\end{aligned}$$

$$= \Pr\left\{\sum_{i=0}^{M-1} TB \log_2(1 + \text{SNR} z_i) < L\right\}. \tag{50}$$

Obviously,  $P_{\text{out}}$  varies with  $L$ . For the Markov model considered in (3),  $L$  bits of *goodput* are removed from the buffer in State 0 with probability  $1 - P_{\text{out}}$  while 0 bit of *goodput* is removed with probability  $P_{\text{out}}$ .

In the following, we first obtain the associated achievable outage effective capacity with given  $L > 0$ . Regarding the Markov modulated processes, we know that [26, Chapter 7, Example 7.2.7]

$$\frac{\Lambda(\theta)}{\theta} = \frac{1}{\theta} \log sp \{ \mathbf{P} \phi(\theta) \} \quad (51)$$

where  $sp\{\cdot\}$  is the spectral radius of the matrix,  $\mathbf{P}$  is the state transition probability matrix (3), and  $\phi(\theta) = \text{diag}\{\phi_0(\theta), \dots, \phi_{M-1}(\theta)\}$  is a diagonal matrix with each component given by the moment generating functions of the *goodput* processes in  $M$  states. With the above characterization of goodput processes in State 0, we have

$$\phi_0(\theta) = P_{\text{out}} + (1 - P_{\text{out}})e^{\theta L}. \quad (52)$$

Note that 0 bit is removed in all other states. Then, we have  $\phi(\theta) = \text{diag}\{P_{\text{out}} + (1 - P_{\text{out}})e^{\theta L}, \underbrace{1, \dots, 1}_{M-1}\}$ .

We are interested in  $-\frac{\Lambda(-\theta)}{\theta}$ . Then, similar to [21, Appendix A], we can show that

$$sp \{ \mathbf{P} \phi(-\theta) \} = p_0 (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L}) y^* \quad (53)$$

where  $y^* > 0$  satisfies

$$\begin{aligned} y^M &= \frac{1 - p_0}{p_0} y^{M-1} + \sum_{m=1}^{M-2} \frac{(1 - p_m) p_{m-1} \cdots p_1}{p_0^m (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})^m} y^{M-1-m} \\ &\quad + \frac{p_{M-2} \cdots p_1}{p_0^{M-1} (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})^{M-1}}. \end{aligned} \quad (54)$$

In addition, we can show that there is only one unique real positive root of  $f(y)$  defined in (14). In this way, we can express the achievable outage effective capacity with given  $L$  as in (12).

Then, the maximum outage effective capacity can be obtained by maximizing over the fixed transmission rate  $L$ . Denote the optimal solution as  $L_{\text{opt}}$ . We can show that there must exist a finite  $L_{\text{opt}}$ . Note that when  $L$  is small, the outage probability approaches 0, and the outage effective capacity is approximately  $\frac{L}{TB}$ , which generally increases with  $L$ . Meanwhile as  $L$  approaches infinity, we know that outage probability approaches 1, i.e., the receiver cannot correctly receive any packet, in which case the outage effective capacity becomes 0. So there must be some finite value of  $L_{\text{opt}}$  that (12) is solved, proving the results in the theorem.  $\square$

### B. Proof of Proposition 1

Note that with the definitions of the parameters, we can see from the discussions in Appendix A that

$$P_{\text{out},m} = p_0 \times p_1 \times \cdots \times p_{m-1} = \prod_{m=0}^{m-1} p_m. \quad (55)$$

By letting (14) equal 0 and multiplying both sides of the resulting equation by  $p_0^M (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})^M$ , we obtain

$$\begin{aligned} & (p_0(P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})y)^M - (1 - p_0)(P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})(p_0(P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})y)^{M-1} \\ & - \sum_{m=1}^{M-2} (1 - p_m)p_{m-1} \cdots p_1 p_0 (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})(p_0(P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})y)^{M-m-1} \\ & - p_{M-2} \cdots p_1 p_0 (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L}) = 0. \end{aligned} \quad (56)$$

Combining (55) with the above equation and letting  $u = p_0(P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})y$ , we have

$$\begin{aligned} & u^M - (P_{\text{out},0} - P_{\text{out},1})(P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})u^{M-1} - \sum_{m=1}^{M-2} (P_{\text{out},m} - P_{\text{out},m+1})(P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L}) \\ & \times u^{M-1-m} - P_{\text{out},M-1}(P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L}) = 0. \end{aligned} \quad (57)$$

First, we can show that as  $\theta \rightarrow 0$ ,  $u \rightarrow 1$ . We know that  $(P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L}) \rightarrow 1$  as  $\theta \rightarrow 0$ .

Then, (57) reduces to

$$u^M - (P_{\text{out},0} - P_{\text{out},1})u^{M-1} - \sum_{m=1}^{M-2} (P_{\text{out},m} - P_{\text{out},m+1})u^{M-1-m} - P_{\text{out},M-1} = 0. \quad (58)$$

It can be easily verified that  $u = 1$  is the unique positive solution to the above equation.

Suppose the Taylor series expansion of  $u$  with respect to small  $\theta$  is given by

$$u = 1 - \zeta\theta + o(\theta), \quad (59)$$

where  $\zeta > 0$  is some constant. According to (13), we can see that

$$\lim_{\theta \rightarrow 0} R_{\text{out}}(\text{SNR}, \theta) = \lim_{\theta \rightarrow 0} -\frac{1}{\theta TB} \log(1 - \zeta\theta + o(\theta)) = \frac{\zeta}{TB}. \quad (60)$$

Therefore, we only need to determine the value of  $\zeta$  to obtain the limit of outage effective



capacity as  $\theta \rightarrow 0$ . The Taylor series expansion of  $P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L}$  with respect to small  $\theta$  is given by

$$P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L} = 1 - (1 - P_{\text{out}})L\theta + o(\theta). \quad (61)$$

Substituting (59) and (61) into (57), and rearranging and combining the coefficients of  $\theta$  gives us

$$\left( M - (M - 1)(P_{\text{out},0} - P_{\text{out},1}) - \sum_{m=1}^{M-2} (P_{\text{out},m} - P_{\text{out},m+1})(M - 1 - m) \right) \zeta = (1 - P_{\text{out}})L, \quad (62)$$

which yields

$$\zeta = \frac{(1 - P_{\text{out}})L}{\sum_{m=0}^{M-1} P_{\text{out},m}}. \quad (63)$$

Combining (60) and (63) and replacing the parameters with the optimal values when  $L = L_{\text{opt}}$  proves the result in the proposition.  $\square$

### C. Proof of Proposition 2

As  $M \rightarrow \infty$ , we know that the outage probability  $P_{\text{out}} \rightarrow 0$ . This is obvious since as  $M \rightarrow \infty$ , retransmissions are triggered as long as the receiver does not receive the packet. We can rewrite (14) as

$$f(y) = y^M - \frac{1 - p_0}{p_0} y^{M-1} - \sum_{i=1}^{M-1} \frac{(1 - p_i)p_{i-1} \cdots p_1}{p_0^i (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})^i} y^{M-1-i} - \frac{\prod_{m=1}^{M-1} p_m}{p_0^{M-1} (P_{\text{out}} + (1 - P_{\text{out}})e^{-\theta L})^{M-1}}. \quad (64)$$

Since the channel is modeled as independently identically distributed (IID) between frames, we have the following characterization for HARQ-T1 protocol

$$p_0 = p_1 = \cdots = p_{M-1} = \Pr \{L > TB \log_2(1 + \text{SNR}z)\} = \Pr \left\{ z < \frac{2^{L/TB} - 1}{\text{SNR}} \right\}. \quad (65)$$

Substituting (65) into (64) , we have

$$\begin{aligned} f(y) &= y^M - \frac{1-p_0}{p_0} \sum_{i=0}^{M-1} (P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L})^{-i} y^{M-1-i} - \frac{1}{(P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L})^{M-1}} \\ &= y^{M-1} \left( y - \frac{1-p_0}{p_0} \sum_{i=0}^{M-1} ((P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L}) y)^{-i} \right) - \frac{1}{(P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L})^{M-1}} \end{aligned} \quad (66)$$

$$= y^{M-1} \left( y - \frac{1-p_0}{p_0} \frac{1 - ((P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L}) y)^{-M}}{1 - ((P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L}) y)^{-1}} \right) - \frac{1}{(P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L})^{M-1}}. \quad (67)$$

Letting  $f(y) = 0$ , we have

$$y - \frac{1-p_0}{p_0} \frac{1 - ((P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L}) y)^{-M}}{1 - ((P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L}) y)^{-1}} = \frac{1}{((P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L}) y)^{M-1}} \quad (68)$$

We can show that  $(P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L}) y > 1$ . Suppose that  $(P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L}) y \leq 1$ . We will have infinite value for the summation in (66) as  $M \rightarrow \infty$ , which in turn returns  $y \rightarrow \infty$  as  $M \rightarrow \infty$  from (68), and as a result  $(P_{\text{out}} + (1-P_{\text{out}})e^{-\theta L}) y \rightarrow \infty$  as  $M \rightarrow \infty$ , violating the assumption.

Taking the limit of both sides of (68) as  $M \rightarrow \infty$  and noting that  $P_{\text{out}} \rightarrow 0$  as  $M \rightarrow \infty$ , we have

$$y - \frac{1-p_0}{p_0} \frac{1}{1 - e^{\theta L}/y} = 0 \quad (69)$$

which after rearrangement yields

$$y^* = \frac{1}{p_0} (1 - p_0 + p_0 e^{\theta L}). \quad (70)$$

Substituting (70) and  $P_{\text{out}} \rightarrow 0$  into (12), we have

$$\lim_{M \rightarrow \infty} R_{\text{out}}(\text{SNR}, \theta) = \max_{L \geq 0} \left\{ -\frac{1}{\theta TB} \log (p_0 + (1-p_0)e^{-\theta L}) \right\} \quad (71)$$

$$= \max_{L \geq 0} \left\{ -\frac{1}{\theta TB} \log \left( 1 - \Pr \left\{ z > \frac{2^{L/TB} - 1}{\text{SNR}} \right\} (1 - e^{-\theta L}) \right) \right\}, \quad (72)$$

proving the results in the proposition.  $\square$

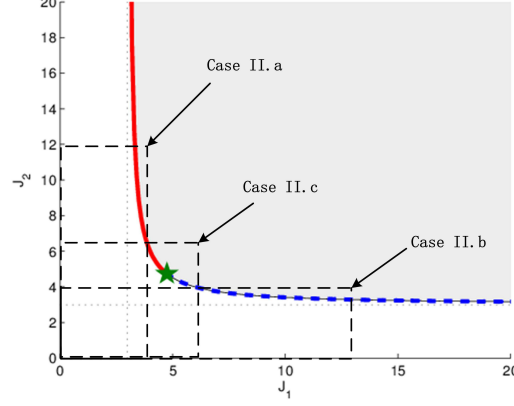


Fig. 8. Illustration of three cases depending on  $(J_{1,\max}, J_{2,\max})$ . The star point denotes  $(J_{th}(\varepsilon), J_{th}(\varepsilon))$ .

#### D. Proof of Theorem 2

The idea of this proof follows that in [23, Appendix D], except that the effective capacity limits as  $\theta \rightarrow \infty$  and  $\theta \rightarrow 0$  are different, and the potential values of  $J_1(\theta_1)$  and  $J_2(\theta_2)$  lie in the range of  $[0, J_{1,\max}]$  and  $[0, J_{2,\max}]$ , respectively. Therefore, when we iterate over all possible  $(J_1, J_2)$  pairs to find the maximum constant arrival rate, we have sliced part of the  $J_1 - J_2$  curve characterized in Lemma 1.

First, we need to check if the statistical delay constraints  $(D_{\max}, \varepsilon)$  can be satisfied. Substituting  $J_1(\theta_1) = J_{1,\max}$  and  $J_2(\theta_2) = J_{2,\max}$  into (20), we can compare the obtained value with  $\varepsilon$ . If the resulting value is larger than  $\varepsilon$ , i.e., **Case I**, the statistical delay constraints cannot be satisfied, and hence the effective capacity is zero. If the resulting value is smaller than  $\varepsilon$ , we can have three different cases depending on the relationship between  $(J_{1,\max}, J_{2,\max})$  and  $(J_{th}(\varepsilon), J_{th}(\varepsilon))$  as shown in Fig. 8. Specifically, we have:

- a) **Case II.a:** If  $J_{1,\max} < J_{th}(\varepsilon)$ , the intersection points of the region  $[0, J_{1,\max}] \times [0, J_{2,\max}]$  with the upper boundary curve  $J_2(\theta_2) = \Phi(J_1(\theta_1))$  lie in the branch with  $J_2 > J_1$ . So we only need to iterate over this branch for potential point  $(J_1, J_2)$  achieving the maximum effective capacity.
- b) **Case II.b:** If  $J_{2,\max} < J_{th}(\varepsilon)$ , we only need to iterate over the branch with  $J_2 < J_1$  for potential point  $(J_1, J_2)$  achieving the maximum effective capacity.
- c) **Case II.c:** Otherwise, we need to consider the two branches jointly to identify the effective

capacity.

**Case II.a:** Assume  $J_{1,\max} < J_{th}(\varepsilon)$ . In this case, we can relieve the statistical delay constraints at the source, i.e., decrease  $J_1(\theta_1)$ , or  $\theta_1$  equivalently. Correspondingly, according to Lemma 1,  $J_2(\theta_2)$ , and hence  $\theta_2$ , should increase. We can show that the queue at the relay will not affect the performance as long as  $\theta_1$  and  $\theta_2$  satisfies the following inequality given by

$$J_1(\theta_1) \leq J_2(\theta_2) + J_1(\theta_1 - \theta_2), \quad (73)$$

and the effective capacity is given by

$$R_E(\theta_1, \theta_2) = \frac{J_1(\theta_1)}{\theta_1}. \quad (74)$$

Note that as  $J_1(\theta_1)$  increases to  $J_{1,\max}$ ,  $\theta_1 \rightarrow \infty$ , and  $J_1(\theta_1 - \theta_2) > 0$ . At the same time,  $J_1(\theta_1) < J_2(\theta_2)$  for this case. The inequality (73) can be satisfied when  $J_1(\theta_1)$  approaches to  $J_{1,\max}$ . On the other hand, as  $J_2(\theta_2)$  increases to  $J_{2,\max}$ ,  $\theta_2 \rightarrow \infty$ , in which case  $\frac{1}{\theta_1 - \theta_2} J_1(\theta_1 - \theta_2)$  approaches to largest possible rate of the first hop [26], i.e.,  $L$  bits/block. Then,  $J_1(\theta_1 - \theta_2)$  approaches minus infinity, and hence the right-hand-side of (73) is less than 0. That is, the inequality (73) cannot be satisfied when  $J_2(\theta_2)$  approaches to  $J_{2,\max}$ . Therefore, there must be a point  $(\overset{\circ}{\theta}_1, \overset{\circ}{\theta}_2) \in \Omega_\varepsilon$  such that  $\overset{\circ}{\theta}_1$  is the smallest value of  $\theta_1$  while (73) can be satisfied with equality at  $(\theta_1, \theta_2)$ . We can show that the effective capacity in this case is given by

$$R_{HARQ-IR}(\varepsilon, D_{\max}, L) = \sup_{(\theta_1, \theta_2) \in \Omega} R_E(\theta_1, \theta_2) = R_E(\overset{\circ}{\theta}_1, \overset{\circ}{\theta}_2) = \frac{J_1(\overset{\circ}{\theta}_1)}{\overset{\circ}{\theta}_1}. \quad (75)$$

Further relieving the statistical delay constraints at the source beyond  $J_1(\overset{\circ}{\theta}_1)$  will result in rate loss since the inequality (73) can not be satisfied, and the queues of the second hop will become the bottle-neck of the system.

**Case II.b:** Assume  $J_{2,\max} < J_{th}(\varepsilon)$ . In this case, we can relieve the statistical delay constraints at the relays, i.e., decrease  $J_2(\theta_2)$ , or  $\theta_2$  equivalently. Correspondingly, according to Lemma 1,  $J_1(\theta_1)$ , and hence  $\theta_1$ , should increase. In this case, we know that the effective capacity is given by

$$\min \left\{ \frac{J_1(\theta_1)}{\theta_1}, \frac{J_2(\theta_2)}{\theta_2} \right\}. \quad (76)$$

Note that as  $J_1(\theta_1)$  increases to  $J_{1,\max}$ ,  $\theta_1 \rightarrow \infty$  and hence  $\frac{J_{1,\max}}{\theta_1}$  approaches to the minimum possible rate of the first hop, which is zero. That is,  $\frac{J_2(\theta_2)}{\theta_2} > \frac{J_1(\theta_1)}{\theta_1}$  as  $J_1(\theta_1) \rightarrow J_{1,\max}$ . Similarly,  $\frac{J_2(\theta_2)}{\theta_2} < \frac{J_1(\theta_1)}{\theta_1}$  as  $J_2(\theta_2) \rightarrow J_{2,\max}$ . Therefore, we can find a unique pair of  $(\check{\theta}_1, \check{\theta}_2) \in \Omega_\varepsilon$  such that  $\frac{J_1(\check{\theta}_1)}{\check{\theta}_1} = \frac{J_2(\check{\theta}_2)}{\check{\theta}_2}$ . We can show that the effective capacity in this case is given by

$$R_{HARQ-IR}(\varepsilon, D_{\max}, L) = \sup_{(\theta_1, \theta_2) \in \Omega} R_E(\theta_1, \theta_2) = R_E(\check{\theta}_1, \check{\theta}_2) = \frac{J_2(\check{\theta}_2)}{\check{\theta}_2}. \quad (77)$$

Further relieving the statistical delay constraints at the relay beyond  $J_2(\check{\theta}_2)$  will result in rate loss since the queues of the first hop will become the bottle-neck of the system.

**Case II.c:** Assume  $J_{1,\max} \geq J_{th}(\varepsilon)$  and  $J_{2,\max} \geq J_{th}(\varepsilon)$ . Now, we need to iterate over two branches with  $J_1 < J_2$  and  $J_1 > J_2$  to find the optimal point  $(J_1, J_2)$  such that the effective capacity can be maximized. Note that this case cover the possibilities in **Case II.a** and **Case II.b**, and also the case in which symmetric delay constraints at the source and relays can achieve the maximum effective capacity, i.e.,  $J_1(\theta_1) = J_2(\theta_2) = J_{th}(\varepsilon)$ .

The details of the derivation for the above claims in (75) and (77) are similar to the proof in [23, Appendix D], and are omitted here. Interested readers are encouraged to find more details in [23, Appendix D].

Meanwhile, we can see that the data correctly received by the destination must also be correctly received at the relays. Therefore, the outage probability of the diamond-relay channels is given by

$$P_{\text{out}} = 1 - (1 - P_{\text{out},s})(1 - P_{\text{out},r}). \quad (78)$$

## REFERENCES

- [1] H. Burton and D. Sullivan, "Errors and error control," *Proc. IEEE*, vol. 60, no. 11, pp. 1293 - 1301, Nov. 1972.
- [2] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1971- 1988, May 2001.
- [3] H. El Gamal, G. Caire, and M. O. Damen, "The MIMO ARQ channel: Diversity-multiplexing-delay tradeoff," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3601 - 3621, Aug. 2006.
- [4] T. V. Chaitanya and E. G. Larsson, "Optimal power allocation for hybrid ARQ with chase combining in IID Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 61, no. 5, pp. 1835 - 1846, May 2013.
- [5] P. Wu and N. Jindal, "Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis," *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1129 - 1141, Apr. 2010.

- [6] T. M. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 5, pp. 572 - 584, Sep. 1979.
- [7] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Trans. Inform. Theory*, vol. 50, no. 12, pp. 3062- 3080, Dec. 2004.
- [8] A. Chelli and M.-S. Alouini, "On the performance of hybrid-ARQ with incremental redundancy and with code combining over relay channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 3860 - 3871, Aug. 2013.
- [9] B. Schein and R. Gallager, "The Gaussian parallel relay network," in Proc. IEEE Int. Symp. Inf. Theory, 2000, p. 22.
- [10] F. Xue and S. Sandhu, "Cooperation in a half-duplex Gaussian diamond relay channels" *IEEE Trans. Inform. Theory*, vol. 53, no. 10, pp. 3806 - 3814, Oct. 2007.
- [11] M. Zamani and A. K. Khandani, "Broadcast approaches to the diamond channel," *IEEE Trans. Inform. Theory*, vol. 60, no. 1, pp. 623 - 643, Jan. 2014.
- [12] F. Parvaresh and R. H. Etkin, "Using superposition codebooks and partial decode-and-forward in low-SNR parallel relay networks," *IEEE Trans. Inform. Theory*, vol. 59, no. 3, pp. 1704-1723, Mar. 2013.
- [13] B. Maham, A. Behnad, and M. Debbah, "Analysis of outage probability and throughput for half-duplex hybrid-ARQ relay channels," *IEEE Trans. Vehi. Technol.*, vol. 61, no. 7, pp. 3061 - 3070, Sep. 2012.
- [14] N. Zlatanov and R. Schober, "Buffer-aided relaying with adaptive link selection-fixed and mixed rate transmission," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 2816-2840, May 2013.
- [15] N. Zlatanov, V. Jamali, and R. Schober, "Achievable rates for the fading half-duplex single relay selection network using buffer-aided relaying," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4494-4507, Aug. 2015.
- [16] D. Wu and R. Negi "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol.2,no. 4, pp. 630-643. July 2003.
- [17] D. Qiao, M. C. Gursoy, and S. Velipasalar, "The impact of QoS constraints on the energy efficiency of fixed-rate wireless transmissions," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5957 - 5969, Dec. 2009.
- [18] N. Gunaseelan, L. Liu, J.-F. Chamberland, and G. H. Huff, "Performance analysis of wireless hybrid-ARQ systems with delay-sensitive traffic," *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1262-1272, Apr. 2010.
- [19] J. Choi, "On large deviations of HARQ incremental redundancy over fading channels," *IEEE Commun. Letters*, vol. 16, no. 6, pp. 913 - 916, June 2012.
- [20] Y. Li, M. C. Gursoy, and S. Velipasalar, "On the throughput of hybrid-ARQ under statistical queuing constraints," *IEEE Trans. Vehi. Technol.*, vol. 64, no. 6, pp. 2725 - 2732, June 2015.
- [21] S. Akin and M. Fidler, "Backlog and delay reasoning in HARQ systems," in 2015 27th International Teletraffic Congress (ITC 2015), pp. 185 - 193, Sep. 2015.
- [22] P. Larsson, J. Gross, H. Al-Zubaidy, L. K. Ramussen, and M. Skoglund, "Effective capacity of retransmission schemes-a recurrence relation approach," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4817-4835, Nov. 2016.
- [23] D. Qiao and M.C. Gursoy, "Statistical delay tradeoffs in buffer-aided two-hop wireless communication systems," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4563-4567, Nov. 2016.
- [24] D. Qiao, "Outage effective capacity of buffer-aided diamond relay systems using HARQ with incremental redundancy," submitted to the 2017 IEEE International Conference on Communications (ICC).
- [25] A. Goldsmith, *Wireless Communications*, 1st ed. Cambridge University Press,
- [26] C.-S. Chang, *Performance Guarantees in Communication Networks*, New York: Springer, 1995.